Bayesian, systems-based methods for analyzing large health data sets

February 7, 2021

Date: 2021-01-30

Version: 1.0

OTKA K 119866

Contents

1 Introduction

Large-scale cohort studies collecting life style, environmental, physiological, clinical and molecular level data, especially genetic information about the participants, provide unprecedented opportunity for a systems-based investigation of genetic, personal, environmental and societal aspects of health, ageing and diseases.

However, as the amount of data is still limited, it is vital to manage uncertainty arising from the combination of knowledge and data in the area of molecular biology, drug discovery, and healthcare. The Bayesian approach provides a principled framework and modern Bayesian methods allow scalable solutions. The completeness (omic-ness) of the data in multiple domains allows the application of novel technologies, such as Bayesian networks (BNs), representing systems of probabilistic dependencies and causal relations.

In earlier works, we investigated the use of BNs in artificial intelligence and machine learning tasks: in knowledge engineering [1], in Bayesian transfer learning [2,3], in explanation generation [4,5], in text-mining [6], in Bayesian feature subset analysis [7,8], in Bayesian effect strength characterization [9], in comparison of general Probabilistic Graphical Models (PGMs) and BNs [OTKA1]. A unique feature of the Bayesian network model class is its inherent ability for inference and learning from a mixture of observations and interventions [10,11]. In an earlier OKTA research (OTKAPD76348), I developed a Bayesian Network-based Bayesian Multilevel Analysis of relevance, dependency, and causal relations [8, 12], which methodology was applied in multiple analyses [13–16], specifically in psychogenetics [17].

Our current OKTA research (K119866) focused on extension, development and application of Bayesian methods to combine knowledge and data in a principled framework. Our main results are as follows:

- *Bayesian multimorbidity maps*: We developed a workflow and adapted our Bayesian inference methods to enhance the construction of Bayesian multimorbidity maps. Using this, we constructed the first epidemilogical multimorbidity map of common diseases [OTKA2], the Bayesian map of the envirome [OTKA3], and maps of multimorbidities related to depression [OTKA1, OTKA4]. Based on our work on Bayesian multimorbidity maps, we initiated and participate in an international research project on depression related multimorbidities TRAJECTOME [OTKA5].

- *Bayesian drug discovery*: We developed a Bayesian drug-target interaction (DTI) prediction method capable for the combination of experimental data and heterogeneous drug and target side information [OTKA6]. We also developed methods to estimate the complex, multivalent binding in protein-protein interactions [OTKA7]. Based on our work on data and knowledge fusion in DTI prediction and its extension towards deep learning methods, we successfully applied for a national grant on informed *de novo* molecule generation DP4D [OTKA8]. Related to our Bayesian, mul-

titask DTI prediction approach, we were invited and participate in an international drug discovery research project MELLODDY [OTKA9].

- *Systems-based fusion in genetics*: We developed methods to investigate multi-trait effects of non-linear gene-gene interactions and gene-environment interactions, especially the effect of modifiable lifestyle factors on multimorbidity and health [OTKA10, OTKA11, OTKA12, OTKA13, OTKA14, OTKA3, OTKA15, OTKA16, OTKA17, OTKA18, OTKA19, OTKA20]. To explore the extension of systems-based fusion using distributed data, we initiated and participated in an international research project HIDUCTION [OTKA21].

We also developed and applied Bayesian systems-based methods for the analysis of clinical laboratory data [OTKA22], and to prepare a multimodal integration, we explored the analysis of health data sets with various modalities, such as from medical imaging and ambient assisted living [OTKA23, OTKA24].

In the final year of our project we also adapted and applied our methods to support national research on SARS-CoV-2/COVID-19.

2 Bayesian multimorbidity maps

Given the rapidly rising prevalence of multimorbidity in modern societies [18–20], its investigation is of primary importance, especially in countries with high multimorbidity rates, such as in Hungary, which has the highest multimorbidity rate in the European Union [21, 22]. Comorbidities, and clusters of multimorbidities became a vital source for the identification common molecular and physiological causal mechanisms [23–29], which can help to identify promising drug candidates targeting all the relevant multimorbidities at once, which is a must to combat polypharmacy, the use instantaneous use of multiple medications for chronic conditions [20]. Indeed, identification of common protective factors relevant to all age-associated diseases is the broadest form of this approach, which could alleviate the burden of aging in modern industrial societies by increasing healthspan [30–36].

Modern large biobanks, such as the UK Biobank data set [37], FinnGen [38], and BioBank Japan [39], provides a joint access to phenotypic and genotypic patient-level data, which allows an unprecedented sample size to explore the shared genetic components of multimorbidities [40–42]

In the project, we focused on the UK Biobank data set. We requested data access in 2013, which were granted by the UK Biobank Application No.1602 for the consortium of the Semmelweis University, Budapest University of Technology and Economics, and The University of Manchester (title: Role of genetics, diet, and comorbidities in depression, principal investigator: Gabriella Juhász). We extended it for the period 2017-2021 and to increase it scope towards general multimorbidities we also requested full access to all the half-million participants to better in 2020. To increase the heterogeneity and sample size in our investigations, we initiated the TRAJECTOME project with the participation of

Finnish, Catalan, and German partners [OTKA5], which allows access to the FinnGen biobank [38] provisionally covering half-million participants and the Health Surveillance System of the region of Catalonia (Spain) with 7.5 million participants [43].

Firstly, the UK Biobank disease codes had to be converted to ICD-10 codes, which ensure compatibility with gene-disease and molecular interactome database levels. We developed and experimented with various approaches to automate the use of the hierarchic descriptors of the absence and presence of the diseases, which led to the construction of the first epidemilogical multimorbidity map of common diseases [OTKA2]. Because of our confirmed interest in environmental effects [OTKA13], we extended the scope of our analysis to include all available environmental effects, including modifiable lifestyle factors, such as diet [OTKA25]. The result of this work led to the construction of the Bayesian map of the envirome [OTKA3]. We also experimented with the use of medication data, which is particularly relevant in causal Bayesian network analysis as interventions [OTKA1, OTKA4]. This work is still in progress, because the recovery of the timing information of the medications exceeded the scope of the project. However, we could access information about disease onsets and we adapted and have already applied our Bayesian Multilevel Analysis method for this data set. The use of temporal information about medication use and disease onset is a natural continuation of our work done in the current project, so we plan to continue our research using these promising resources [24, 41].

We also adapted and applied Bayesian network based methods for the analysis of clinical laboratory data [OTKA22].

3 Bayesian drug discovery

In multimorbidity research, we started to explore the applicability of information about medication in the UK Biobank, but large-scale, comprehensive data sets about real-world drug effects are also novel information sources in drug discovery. Earlier we developed drug repositioning methods focusing on the fusion of heterogeneous information sources, such as chemical, target, and side-effect similarities [44–46]. In the current project, we extended our earlier scope to drug candidates and developed methods for the early drug discovery phase utilizing novel large-scale, cross-domain linked open data [47–49] and large-scale drug-target interaction bioactivity data sets [50, 51].

We developed a Bayesian drug-target interaction (DTI) prediction method (VB-MK-LMF) capable for the combination of experimental data and heterogeneous side information about drug and target drugs and targets [OTKA6]. The developed Bayesian, multitask VB-MK-LMF method provides a principled framework for the fusion of large-scale information about chemical compounds, binding sites, drug targets, protein-protein interactions, and gene regulatory networks, which is currently extended in an international drug discovery research project MELLODDY [OTKA9]. The MELLODDY consortium contains leading international pharmaceutical companies with significantly different pri-

vate data sets from the public data set [52–55], which hopefully will lead to novel theoretical extensions of our research, e.g., related to deep learning [56].

The multitask nature of the VB-MK-LMF method makes it an ideal candidate for multitarget drug discovery [57]. It also allows the incorporation of information about drug targets, protein-protein interactions, gene regulatory networks, which are essential in polypharmacology [58], and even information about shared genetics in multimorbidities, which helps to cope with polypharmacy [20]. To extend our research towards synthetically accessible drug candidates without any empirical bioactivity data, we successfully applied for a national grant on informed *de novo* molecule generation DP4D [OTKA8].

In addition to predicting interactions between small compounds and protein targets, we also developed methods to estimate the complex, multivalent binding in protein-protein interactions [OTKA7]. Currently, we are extending the model to capture the sequential, competitive nature of multivalent binding, which requires efficient combination of combinatorial search and Markov Chain Monte Carlo sampling.

4 Systems-based fusion in genetics

The increasing sample size in genome-wide association studies (GWASs) provides a lower and lower upper bound on main effects of common genetic variants, confirming the long-anticipated genetic architecture of high-number of genetic variants with infinitesimally small effects, but it also suggests the presence of gene-gene, gene-environment interactions and conditional relevance of genetic variants in common diseases, such as in depression [OTKA13, OTKA14]. Standard, set-based enrichment methods allows the aggregate analysis of the effects of variants in a given set corresponding to any functional aspect, e.g., at gene and pathway levels, which also takes into account the pairwise dependencies of variants [59–61]. Standard enrichment methods using GWAS summary statistics became very popular, but the availability of large-scale biobank data allows multiple extensions: (1) testing directly the significance of a set of variants using mixed models [62–72]; (2) propagating gene level evidences in context-specific gene-gene networks [73–80]; and (3) combining evidences from multiple structured traits and multimorbidities [23–29]. We developed a comprehensive workflow to support these options: (1) we implemented a GPU-based generalized linear mixed model, which can directly evaluate tests for genes and pathways; (2) we investigated multiple protein-protein networks, gene regulatory networks, and molecular networks related to depression and network propagation settings [OTKA16]; and (3) we evaluated efficient methods to estimate genetic correlations [81–83]. To adjust for the confounding effects of shared genetic background, we integrated their multivariate extensions into our Bayesian multimorbidity learning method, which can also take into account environmental conditions [OTKA2, OTKA13, OTKA25, OTKA3]. Thus, in addition to the confounding effects of environmental factors and medications, confounding effects of shared genetic factors can be also decomposed and filtered from the

multimorbidity maps [84].

We applied these methods to explore the genetic background of allergy [OTKA10], depression and its related multimorbidities research [OTKA15, OTKA26], education and intelligence [OTKA18], and healthspan [OTKA17,OTKA19,OTKA20]. Currently, we are investigating these methods to explore the joint effects of modifiable lifestyle factors, such as diet and exercise on depression related multimorbidities [OTKA26].

To explore the extension of these network-based fusion methods using distributed data sets in personalized medicine, we initiated and participated in an international research project HIDUCTION [OTKA21].

5 Multimorbidity analysis and repositioning in COVID-19

In the final year of our project we also adapted and applied our methods to support national research on SARS-CoV-2/COVID-19.

The VB-MK-LMF method can be also applied in drug discovery and repositioning against SARS-CoV-2/COVID-19 using phenotypic screening. Thus, we adapted our drug discovery methods to support the drug repositioning efforts of the Repositioning workgroup led by P. Mátyus in Hungary's Coronavirus research action group.

We also applied our Bayesian multimorbidity analysis of COVID-19 deceased in 2020 in Hungary [OTKA27], which were extended using a representative Hungarian biobank. Currently, we are synthesizing these results to construct multimorbidity-based risk groups for the early detection of severe COVID-19 cases and to support the design of vaccination policy [OTKA28].

6 OTKA References

[OTKA1] Gabor Hullam, Gabriella Juhasz, Bill Deakin, and Peter Antal. Structural and parametric uncertainties in full bayesian and graphical lasso based approaches: Beyond edge weights in psychological networks. In *2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE, 2017.

[OTKA2] Peter Marx, Peter Antal, Bence Bolgar, Gyorgy Bagdy, Bill Deakin, and Gabriella Juhasz. Comorbidities in the diseasome are more apparent than real: What bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, 13(6):e1005487, 2017.

[OTKA3] Gabor Hullam, Peter Antal, Peter Petschner, Xenia Gonda, Gyorgy Bagdy, Bill Deakin, and Gabriella Juhasz. The UKB envirome of depression: from interactions to synergistic effects. *Scientific reports*, 9(1):1–19, 2019.

[OTKA4] M. Vetró, G. Hullám, G. Juhász, and P. Antal. A depresszió környezeti faktorainak vizsgálata oksági elemzési módszerekkel. In *Orvosi Informatika 2020 – XXXIII. Neumann Kollokvium*, pages 106–112, 2020.

[OTKA5] The TRAJECTOME Consortium. Temporal disease map based stratification of depression-related multimorbidities: towards quantitative investigations of patient trajectories and predictions of multi-target drug candidates. https://semmelweis.hu/trajectome/en/, 2021-2023. ERA PerMed Joint Translational Call 2019 (No. 2019-2.1.7-ERA-NET-2020-00005).

[OTKA6] Bence Bolgár and Péter Antal. VB-MK-LMF: fusion of drugs, targets and interactions using variational bayesian multiple kernel logistic matrix factorization. *BMC bioinformatics*, 18(1):440, 2017.

[OTKA7] Wesley J Errington, Bence Bruncsics, and Casim A Sarkar. Mechanisms of noncanonical binding dynamics in multivalent protein–protein interactions. *Proceedings of the National Academy of Sciences*, 116(51):25659–25667, 2019.

[OTKA8] P. Antal. Deep priors for drugs (dp4d): De novo hatóanyagjelölt generálás nagy mennyiségű bioaktivitási információkat felhasználó mély megerősítéses tanulással: több szempont együttes optimalizálása mesterséges intelligenciával a korai gyógyszerkutatásban, 2020. Richter Témapályázat (TP13-2019-TP13/017).

[OTKA9] The MELLODDY Consortium. Machine learning ledger orchestration for drug discovery. https://www.melloddy.eu/, 2019-2022. H2020/IMI2 (G.A.No.: 831472).

[OTKA10] Viktor Molnár, Adrienne Nagy, Lilla Tamási, Gabriella Gálffy, Renáta Böcskei, András Bikov, Ibolya Czaller, Zsuzsanna Csoma, Magdolna Krasznai, Csilla Csáki, et al. From genomes to diaries: A 3-year prospective, real-life study of ragweed-specific sublingual immunotherapy. *Immunotherapy*, 9(15):1279–1294, 2017.

[OTKA11] Lili E Fodor, András Gézsi, et al. Investigation of the possible role of the hippo/yap1 pathway in asthma and allergy. *Allergy, asthma & immunology research*, 9(3):247, 2017.

[OTKA12] Gabriella Juhász, E Csepany, Máté Magyar, Andrea Edit Édes, Nóra Eszlári, Gabor Hullam, Péter Antal, Gy Kokonyei, IM Anderson, JFW Deakin, et al. Variants in the cnr1 gene predispose to headache with nausea in the presence of life stress. *Genes, Brain and Behavior*, 16(3):384–393, 2017.

[OTKA13] Xenia Gonda, Gabor Hullam, Peter Antal, Nora Eszlari, Peter Petschner, Tomas GM Hökfelt, Ian Muir Anderson, John Francis William Deakin, Gabriella Juhasz, and Gyorgy Bagdy. Significance of risk polymorphisms for depression depends on stress exposure. *Scientific reports*, 8(1):1–10, 2018.

[OTKA14] Xenia Gonda, Peter Petschner, Nora Eszlari, Daniel Baksa, Andrea Edes, Peter Antal, Gabriella Juhasz, and Gyorgy Bagdy. Genetic variants in major depressive disorder: From pathophysiology to therapy. *Pharmacology & therapeutics*, 194:22–43, 2019.

[OTKA15] Nora Eszlari, Andras Millinghoffer, Peter Petschner, Xenia Gonda, Daniel Baksa, Attila J Pulay, János M Réthelyi, Gerome Breen, John Francis William Deakin, Peter Antal, et al. Genome-wide association analysis reveals kctd12 and mir-383-binding genes in the background of rumination. *Translational psychiatry*, 9(1):1–12, 2019.

[OTKA16] Bence Bruncsics and Peter Antal. A multi-trait evaluation of network propagation for gwas results. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–6. IEEE, 2019.

[OTKA17] Georg Fuellen, Ludger Jansen, Alan A Cohen, Walter Luyten, Manfred Gogol, Andreas Simm, Nadine Saul, Francesca Cirulli, Alessandra Berry, Peter Antal, et al. Health and aging: unifying concepts, scores, biomarkers and pathways. *Aging and disease*, 10(4):883, 2019.

[OTKA18] Peter Przemyslaw Ujma, Nóra Eszlári, Andras Millinghoffer, Bence Bruncsics, Peter Petschner, Péter Antal, Bill Deakin, Gerome Breen, Gyorgy Bagdy, and Gabriella Juhasz. Genetic effects on educational attainment in hungary. *bioRxiv*, 2020.

[OTKA19] Steffen Möller, Nadine Saul, Alan A Cohen, Rüdiger Köhling, Sina Sender, Hugo Murua Escobar, Christian Junghanss, Francesca Cirulli, Alessandra Berry, Peter Antal, et al. Healthspan pathway maps in c. elegans and humans highlight transcription, proliferation/biosynthesis and lipids. *Aging (Albany NY)*, 12(13):12534, 2020.

[OTKA20] V. Várhegyi, V. Molnár, Sárközy Gézsi, A, Antal P. P., and M.J. Molnár. Magyar genomikai egészségtárház az egészséges hosszú élet kutatásának szolgálatában. *Orvosi hetilap.* accepted.

[OTKA21] The HIDUCTION Consortium. Privacy preserving data and knowledge fusion in personalized biomedicine. https://celsalliance.eu/selected%20projects, 2017-2019. Central Europe Leuven Strategic Alliance.

[OTKA22] Zeyneb Guenfoud and Péter Antal. Bayesian exploration of dependencies of laboratory tests and evaluation of test redundancy. In

2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS), pages 1–6. IEEE, 2018.

[OTKA23] Béla Pataki, Enikő Sirály, and György Strausz. Improving estimation of mental wellness using computer games. In *European Medical and Biological Engineering Conference*, pages 655–663. Springer, 2020.

[OTKA24] Dániel Hadházi and Gábor Horváth. Anisotropic iteratively reweighted TV regularized reconstruction for linear tomosynthesis. In *European Medical and Biological Engineering Conference*, pages 84–94. Springer, 2020.

[OTKA25] G Juhasz, P Petschner, G Bagdy, B Deakin, P Antal, and G Hullam. Contributing factors in the comorbidity of depression and pain: A bayesian approach. *European Neuropsychopharmacology*, 29:S290–S291, 2019.

[OTKA26] Nóra Eszlári, Bence Bruncsics, Andras Millinghoffer, Hullam Gabor, , Peter Petschner, Gonda Xenia, Gerome Breen, Péter Antal, Gyorgy Bagdy, Bill Deakin, and Gabriella Juhasz. Biology of perseverative negative thinking: the role of timing and folate intake, 2020. submitted.

[OTKA27] T. Nagy, B. Bruncsics, and P. Antal. Bayesian network multimorbidity models in covid-19 mortality. In *28th Minisymposium of the Department of Measurement and Information Systems*, pages 1–6, 2021.

[OTKA28] T. Nagy, Zs Bagyura, B. Bruncsics, G. Juhasz, B. Merkely, and P. Antal. Systems-based analysis of multimorbidities of covid-19 deceased in 2020 in hungary. *Orvosi hetilap*. in preparation.

7 References

[1] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote. Bayesian networks in ovarian cancer diagnosis: Potential and limitations. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med. Sys. (CBMS-2000)*, pages 103–109, 2000.

[2] P. Antal, G. Fannes, H. Verrelst, B. De Moor, and J. Vandewalle. Incorporation of prior knowledge in black-box models: Comparison of transformation methods from Bayesian network to multilayer perceptrons. In *Workshop on Fusion of Domain Knowledge with Data for Decision Support, 16th Uncertainty in Artificial Intelligence Conference*, pages 42–48, 2000.

[3] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial Intelligence in Medicine*, 29:39–60, 2003.

[4] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Annotated Bayesian networks: a tool to integrate textual and probabilistic medical knowledge. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med. Sys. (CBMS-2001)*, pages 177–182, 2001.

[5] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Domain knowledge based information retrieval language: an application of annotated Bayesian networks in ovarian cancer domain. In *Proc. of the 15th IEEE Symp. on Computer-Based Medical Sys. (CBMS-2002)*, pages 213–218, 2002.

[6] P. Antal, P. Glenisson, G. Fannes, J. Mathijs, Y. Moreau, and B. De Moor. On the potential of domain literature for clustering and Bayesian network learning. In *Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (ACM-KDD-2002)*, pages 405–414, 2002.

[7] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. De Moor. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30:257–281, 2004.

[8] P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus. A Bayesian multilevel analysis of feature relevance. In *Workshop on New challenges for feature selection in data mining and knowledge discovery (FSDM 2008) at The 19th European Conference on Machine Learning (ECML 2008)*, 2008.

[9] Gabor Hullam and Peter Antal. Estimation of effect size posterior using model averaging over bayesian network structures and parameters. In *Proceedings of the Sixth European Workshop on Probabilistic Graphical Models. Granada, Spain*, pages 147–154, 2012.

[10] G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence (UAI-1999)*, pages 116–125. Morgan Kaufmann, 1999.

[11] C. Glymour and G. F. Cooper. *Computation, Causation, and Discovery.* AAAI Press, 1999.

[12] P Antal, A Millinghoffer, G Hullám, G Hajós, P Sárközy, A Gézsi, C Szalai, and A Falus. Bayesian, systems-based, multilevel analysis of associations for complex phenotypes: from interpretation to decisions. *Probabilistic Graphical Models for Genetics, Genomics and Postgenomics. Oxford University Press: Oxford, UK*, 2014.

[13] Ildikó Ungvári, Gábor Hullám, Péter Antal, Petra Sz Kiszel, András Gézsi, Éva Hadadi, Viktor Virág, Gergely Hajós, András Millinghoffer, Adrienne Nagy, et al. Evaluation of a partial genome screening of two asthma susceptibility regions using bayesian network based bayesian multilevel analysis of relevance. *PLoS One*, 7(3):e33573, 2012.

[14] Andrea Vereczkei, Zsolt Demetrovics, Anna Szekely, Peter Sarkozy, Peter Antal, Agnes Szilagyi, Maria Sasvari-Szekely, and Csaba Barta. Multivariate analysis of dopaminergic gene variants as risk factors of heroin dependence. *PLoS One*, 8(6):e66592, 2013.

[15] Gabor Varga, Anna Szekely, Peter Antal, Peter Sarkozy, Zsofia Nemoda, Zsolt Demetrovics, and Maria Sasvari-Szekely. Additive effects of serotonergic and dopaminergic polymorphisms on trait impulsivity. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159(3):281–288, 2012.

[16] Orsolya Lautner-Csorba, András Gézsi, Ágnes F Semsei, Péter Antal, Dániel J Erdélyi, Géza Schermann, Nóra Kutszegi, Katalin Csordás, Márta Hegyi, Gábor Kovács, et al. Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by bayesian network based bayesian multilevel analysis of relevance. *BMC medical genomics*, 5(1):42, 2012.

[17] Gabriella Juhasz, Gabor Hullam, Nora Eszlari, Xenia Gonda, Peter Antal, Ian Muir Anderson, Tomas GM Hökfelt, JF William Deakin, and Gyorgy Bagdy. Brain galanin system genes interact with life stresses in depression-related phenotypes. *Proceedings of the National Academy of Sciences*, 111(16):E1666–E1673, 2014.

[18] Anna J Koné Pefoyo, Susan E Bronskill, Andrea Gruneir, Andrew Calzavara, Kednapa Thavorn, Yelena Petrosyan, Colleen J Maxwell, YuQing Bai, and Walter P Wodchis. The increasing burden and complexity of multimorbidity. *BMC public health*, 15(1):415, 2015.

[19] Andrew Kingston, Louise Robinson, Heather Booth, Martin Knapp, Carol Jagger, and MODEM project. Projections of multi-morbidity in the older population in england to 2035: estimates from the population ageing and care simulation (pacsim) model. *Age and ageing*, 47(3):374–380, 2018.

[20] Bruce Guthrie, Boikanyo Makubate, Virginia Hernandez-Santiago, and Tobias Dreischulte. The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995–2010. *BMC medicine*, 13(1):74, 2015.

[21] Sara Afshar, Paul J Roderick, Paul Kowal, Borislav D Dimitrov, and Allan G Hill. Multimorbidity and the inequalities of global ageing: a cross-sectional study of 28 countries using the world health surveys. *BMC Public Health*, 15(1):776, 2015.

[22] Raffaele Palladino, John Tayu Lee, Mark Ashworth, Maria Triassi, and Christopher Millett. Associations between multimorbidity, healthcare utilisation and health status: evidence from 16 european countries. *Age and ageing*, 45(3):431–435, 2016.

[23] Michael Krauthammer, Charles A Kaufmann, T Conrad Gilliam, and Andrey Rzhetsky. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in alzheimer's disease. *Proceedings of the National Academy of Sciences*, 101(42):15148–15153, 2004.

[24] Andrey Rzhetsky, David Wajngurt, Naeun Park, and Tian Zheng. Probing genetic overlap among complex human phenotypes. *Proceedings of the National Academy of Sciences*, 104(28):11694–11699, 2007.

[25] Kanix Wang, Hallie Gaitsch, Hoifung Poon, Nancy J Cox, and Andrey Rzhetsky. Classification of common human diseases derived from shared genetic and environmental determinants. *Nature genetics*, 49(9):1319, 2017.

[26] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236, 2015.

[27] Oriol Canela-Xandri, Konrad Rawlik, and Albert Tenesa. An atlas of genetic associations in uk biobank. *Nature genetics*, 50(11):1593–1599, 2018.

[28] Tom G Richardson, Sean Harrison, Gibran Hemani, and George Davey Smith. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife*, 8:e43657, 2019.

[29] Adrian Cortes, Patrick K Albers, Calliope A Dendrou, Lars Fugger, and Gil McVean. Identifying cross-disease components of genetic risk across hospital data in the uk biobank. *Nature genetics*, 52(1):126–134, 2020.

[30] James F Fries. Aging, natural death, and the compression of morbidity. *New England Journal of Medicine*, 303(3):130–135, 1980.

[31] James F Fries, Bonnie Bruce, and Eliza Chakravarty. Compression of morbidity 1980–2011: a focused review of paradigms and progress. *Journal of aging research*, 2011, 2011.

[32] Eileen M Crimmins and Hiram Beltrán-Sánchez. Mortality and morbidity trends: is there compression of morbidity? *The Journals of Gerontology: Series B*, 66(1):75–86, 2011.

[33] Eileen M Crimmins. Lifespan and healthspan: past, present, and promise. *The Gerontologist*, 55(6):901–911, 2015.

[34] Malene Hansen, Ao-Lin Hsu, Andrew Dillin, and Cynthia Kenyon. New genes tied to endocrine, metabolic, and dietary regulation of lifespan from a caenorhabditis elegans genomic rnai screen. *PLoS genetics*, 1(1):e17, 2005.

[35] Andrew R Harper, Shalini Nayee, and Eric J Topol. Protective alleles and modifier variants in human health and disease. *Nature Reviews Genetics*, 16(12):689, 2015.

[36] Galina A Erikson, Dale L Bodian, Manuel Rueda, Bhuvan Molparia, Erick R Scott, Ashley A Scott-Van Zeeland, Sarah E Topol, Nathan E Wineinger, John E Niederhuber, Eric J Topol, et al. Whole-genome sequencing of a healthy aging cohort. *Cell*, 165(4):1002–1011, 2016.

[37] Naomi Allen, Cathie Sudlow, Paul Downey, Tim Peakman, John Danesh, Paul Elliott, John Gallacher, Jane Green, Paul Matthews, Jill Pell, et al. Uk biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1(3):123–126, 2012.

[38] Iiro Hämäläinen, Outi Toernwall, Birgit Simell, Kurt Zatloukal, Markus Perola, and Gert-Jan B van Ommen. Role of academic biobanks in public–private partnerships in the european biobanking and biomolecular resources research infrastructure community. *Biopreservation and biobanking*, 17(1):46–51, 2019.

[39] Saori Sakaue, Masahiro Kanai, Yosuke Tanigawa, Juha Karjalainen, Mitja Kurki, Seizo Koshiba, Akira Narita, Takahiro Konuma, Kenichi Yamamoto, Masato Akiyama, et al. A global atlas of genetic associations of 220 deep phenotypes. *medRxiv*, 2020.

[40] Karen A Mather and Anbupalam Thalamuthu. Unraveling the genetic contributions to complex traits across different ethnic groups. *Nature medicine*, 26(4):467–469, 2020.

[41] Yen-Chen A Feng, Tian Ge, Mattia Cordioli, Andrea Ganna, Jordan W Smoller, Benjamin M Neale, et al. Findings and insights from the genetic investigation of age of first reported occurrence for complex disorders in the uk biobank and finngen. *medRxiv*, 2020.

[42] Saori Sakaue, Masahiro Kanai, Juha Karjalainen, Masato Akiyama, Mitja Kurki, Nana Matoba, Atsushi Takahashi, Makoto Hirata, Michiaki Kubo, Koichi Matsuda, et al. Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan. *Nature Medicine*, 26(4):542–548, 2020.

[43] Emili Vela, Ákos Tényi, Isaac Cano, David Monterde, Montserrat Cleries, Anna Garcia-Altes, Carme Hernandez, Joan Escarrabill, and Josep Roca. Population-based analysis of patients with copd in catalonia: a cohort study with implications for clinical management. *BMJ open*, 8(3):e017283, 2018.

[44] A Arany, Bence Bolgár, Balázs Balogh, Peter Antal, and Péter Mátyus. Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources. *Current medicinal chemistry*, 20(1):95–107, 2013.

[45] Bence Bolgár, Adam Arany, Gergely Temesi, Balázs Balogh, Péter Antal, and Peter Matyus. Drug repositioning for treatment of movement disorders: from serendipity to rational discovery strategies. *Current topics in medicinal chemistry*, 13(18):2337–2363, 2013.

[46] Gergely Temesi, Bence Bolgár, Ádám Arany, Csaba Szalai, Péter Antal, and Péter Mátyus. Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy. *Future medicinal chemistry*, 6(5):563–575, 2014.

[47] Antony J. Williams, Lee Harland, Paul Groth, Stephen Pettifer, Christine Chichester, Egon L. Willighagen, Chris T. Evelo, Niklas Blomberg, Gerhard Ecker, Carole Goble, and Barend Mons. Open PHACTS: Semantic interoperability for drug discovery. *Drug Discovery Today*, 17(21-22):1188–1198, 2012.

[48] Kamal Azzaoui, Edgar Jacoby, Stefan Senger, Emiliano Cuadrado Rodríguez, Mabel Loza, Barbara Zdrazil, Marta Pinto, Antony J. Williams, Victor De La Torre, Jordi Mestres, Manuel Pastor, Olivier Taboureau, Matthias Rarey, Christine Chichester, Steve Pettifer, Niklas Blomberg, Lee Harland, Bryn Williams-Jones, and Gerhard F. Ecker. Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discovery Today*, 18(17-18):843–852, 2013.

[49] Daria Goldmann, Floriane Montanari, Lars Richter, Barbara Zdrazil, and Gerhard F Ecker. Exploiting open data: a new era in pharmacoinformatics. *Future medicinal chemistry*, 6(5):503–514, 2014.

[50] Anna Gaulton, Louisa J. Bellis, a. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):1100–1107, 2012.

[51] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, et al. Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of cheminformatics*, 9(1):1–9, 2017.

[52] Christopher Southan, Péter Várkonyi, and Sorel Muresan. Complementarity between public and commercial databases: new opportunities in medicinal chemistry informatics. *Current Topics in Medicinal Chemistry*, 7(15):1502–1508, 2007.

[53] Christopher Southan, Péter Vrkonyi, and Sorel Muresan. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *Journal of Cheminformatics*, 1(1):1–17, 2009.

[54] Carole Goble, Alasdair J G Gray, Lee Harland, Karen Karapetyan, Antonis Loizou, Ivan Mikhailov, Yrjänä Rankka, Stefan Senger, Valery Tkachenko, Antony Williams, and Egon Willighagen. Incorporating Private and Commercial Data into an Open Linked Data Platform for Drug Discovery. *12th International Semantic Web Conference (ISWC)*, pages 1–16, 2013.

[55] Christopher A Lipinski, Nadia K Litterman, Christopher Southan, Antony J Williams, Alex M Clark, and Sean Ekins. Parallel worlds of public and commercial bioactive chemistry data: Miniperspective. *Journal of medicinal chemistry*, 58(5):2068, 2015.

[56] Noé Sturm, Andreas Mayr, Thanh Le Van, Vladimir Chupakhin, Hugo Ceulemans, Joerg Wegner, Jose-Felipe Golib-Dzib, Nina Jeliazkova, Yves Vandriessche, Stanislav Böhm, et al. Industry-scale application and evaluation of deep learning for drug target prediction. *Journal of Cheminformatics*, 12:1–13, 2020.

[57] Péter Mátyus. Több támadáspontú gyógyszerek: múlt, jelen és jövő= multi-targeting drugs: past, present and future. *Orvosi hetilap*, 161(14):523–531, 2020.

[58] Peter Csermely, Tamás Korcsmáros, Huba JM Kiss, Gábor London, and Ruth Nussinov. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & therapeutics*, 138(3):333–408, 2013.

[59] Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, 2010.

[60] David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS computational biology*, 12(1):e1004714, 2016.

[61] Ryan Sun, Shirley Hui, Gary D Bader, Xihong Lin, and Peter Kraft. Powerful gene set analysis in gwas with the generalized berk-jones statistic. *PLoS genetics*, 15(3):e1007530, 2019.

[62] BJ Hayes, ME Goddard, et al. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

[63] Mike Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136(2):245–257, 2009.

[64] Ben John Hayes, Peter M Visscher, and Michael E Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research*, 91(1):47–60, 2009.

[65] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203, 2006.

[66] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459, 2010.

[67] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[68] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833, 2011.

[69] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nature methods*, 9(6):525, 2012.

[70] Jennifer Listgarten, Christoph Lippert, and David Heckerman. Fast-lmm-select for addressing confounding from spatial structure and rare variants. *Nature Genetics*, 45(5):470, 2013.

[71] Christian Widmer, Christoph Lippert, Omer Weissbrod, Nicolo Fusi, Carl Kadie, Robert Davidson, Jennifer Listgarten, and David Heckerman. Further improvements to linear mixed models for genome-wide association studies. *Scientific reports*, 4:6874, 2014.

[72] Rachel Moore, Francesco Paolo Casale, Marc Jan Bonder, Danilo Horta, Lude Franke, Inês Barroso, and Oliver Stegle. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature genetics*, 51(1):180–186, 2019.

[73] Trey Ideker, Janusz Dutkowski, and Leroy Hood. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, 144(6):860–863, 2011.

[74] Mark DM Leiserson, Jonathan V Eldridge, Sohini Ramachandran, and Benjamin J Raphael. Network analysis of gwas data. *Current opinion in genetics & development*, 23(6):602–610, 2013.

[75] Mark DM Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, 47(2):106, 2015.

[76] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551, 2017.

[77] M.D.M. Leiserson, F. Vandin, H-T. Wu, J.R Dobson, J.V. Eldridge, J.L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M.S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G.A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B.J. Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, 2015.

[78] Priyanka Nakka, Benjamin J Raphael, and Sohini Ramachandran. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics*, 204(2):783–798, 2016.

[79] Matthew A Reyna, Mark DM Leiserson, and Benjamin J Raphael. Hierarchical hotnet: identifying hierarchies of altered subnetworks. *Bioinformatics*, 34(17):i972–i980, 2018.

[80] Oded Magger, Yedael Y Waldman, Eytan Ruppin, and Roded Sharan. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS computational biology*, 8(9):e1002690, 2012.

[81] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291, 2015.

[82] Doug Speed, John Holmes, and David J Balding. Evaluating and improving heritability models using summary statistics. *Nature Genetics*, 52(4):458–462, 2020.

[83] Zheng Ning, Yudi Pawitan, and Xia Shen. High-definition likelihood inference of genetic correlations across human complex traits. *Nature genetics*, 52(8):859–864, 2020.

[84] Verneri Anttila, Brendan Bulik-Sullivan, Hilary K Finucane, Raymond K Walters, Jose Bras, Laramie Duncan, Valentina Escott-Price, Guido J Falcone, Padhraig Gormley, Rainer Malik, et al. Analysis of shared heritability in common disorders of the brain. *Science*, 360(6395):eaap8757, 2018.